

Pam:
31:
(*58)
HEL

REPRINT SERIES NO. 29

CENSUS STATISTICS THROUGH COMBINED USE OF SURVEYS AND REGISTERS?

BY

J. HELDAL, A. RYGH SWENSEN AND I. THOMSEN

REPRINT FROM STATISTICAL JOURNAL OF THE UNITED NATIONS ECE 5 (1987)

NORTH-HOLLAND

CENTRAL BUREAU OF STATISTICS

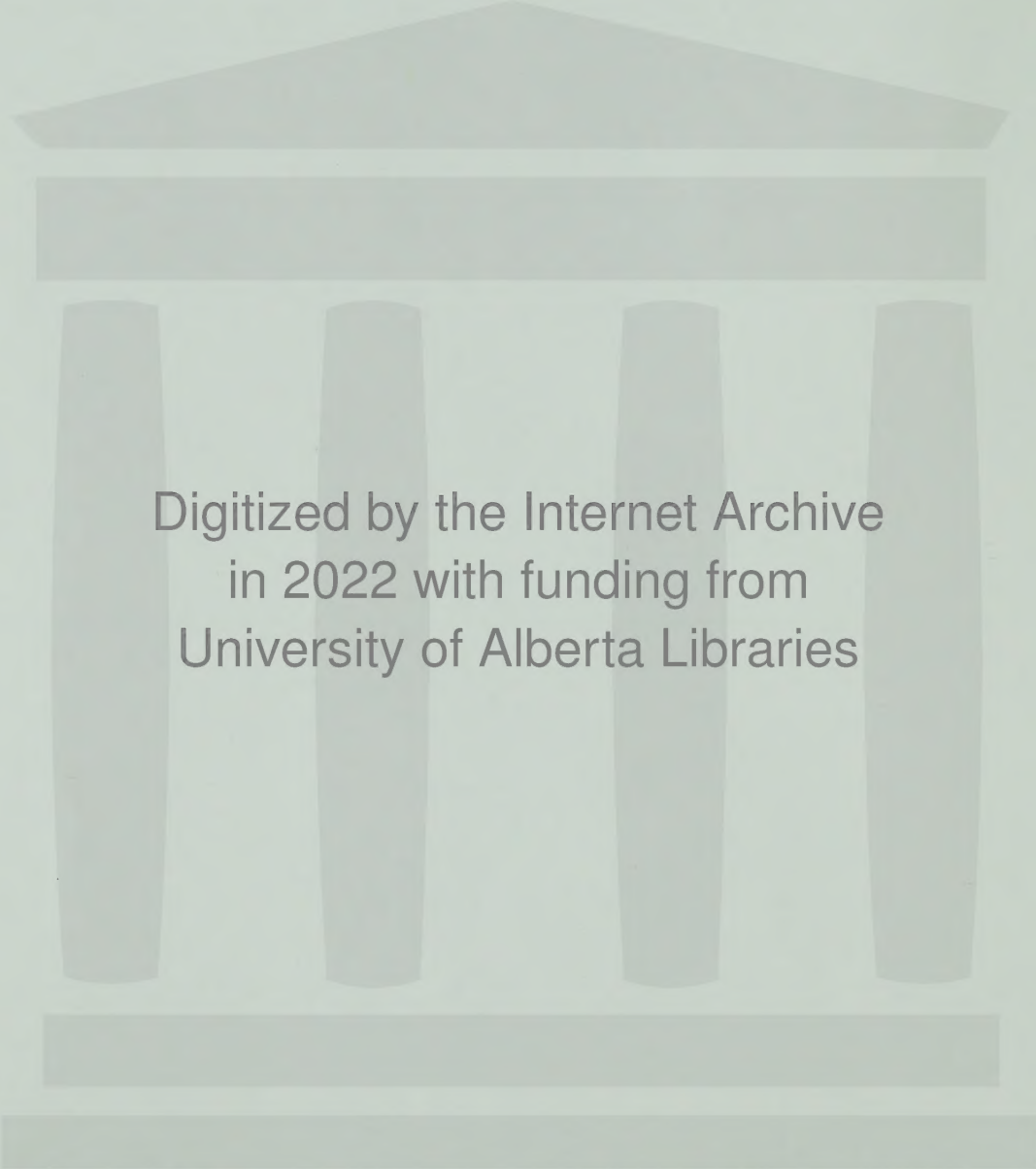
OSLO - NORWAY

1988

ISBN 82-537-2639-2
ISSN 0800-7500

POLAR
PAM
277

POLAR-PAM



Digitized by the Internet Archive
in 2022 with funding from
University of Alberta Libraries

<https://archive.org/details/censusstatistics00held>

Census statistics through combined use of surveys and registers?

J. Heldal, A. Rygh Swensen and I. Thomsen

At the Central Bureau of Statistics of Norway the feasibility of producing census statistics through combined use of surveys and registers has been studied since 1985. Different methods for combining information from registers and surveys have been investigated in order to produce census statistics of satisfactory quality.

1. Introduction and background

In Norway, the first population census was taken more than 200 years ago. During the major part of this period, the census has been considered to be one of the main sources of social and demographic statistics for planning and research. For the last twenty years though, a large amount of the information, which was previously collected through the census, is available from administrative registers. At the same time, the costs of a census have not decreased substantially.

Against this background, it is not surprising that statisticians try to find new methods in order to replace the traditional population census by less expensive data collection methods. Due to the long history and high prestige of the census, however, it is equally understandable that these efforts are met with a good deal of scepticism by some statisticians and most users of census data.

In this paper we present some results from a study on the feasibility of producing census statistics through the combined use of surveys and registers. The main finding is that if one is willing to relax a little on the need for accuracy of statistics for "small" areas, the large majority of the census statistics can be produced by such a combined use of data sources.

Ever since the establishment of the first registers, the Norwegian statistical office has made considerable efforts to influence the content and maintenance of

Johan Heldal is a mathematical statistician working as a senior executive in the Methods Unit of the Central Bureau of Statistics of Norway where he is employed since eight years ago. His main assignments are quality control and combined use of sample surveys, censuses and registers.

Anders Rygh Swensen was born 1949. He graduated in statistics from University of Oslo 1975 and got his Ph.D. from University of California, Berkeley 1980. During the period 1981-83 he was employed as a research associate at the Norwegian Telecommunications Administration. Since 1984 he is with the Methods Unit, Central Bureau of Statistics of Norway.

Ib Thomsen was born 1939. He graduated in statistics from University of Oslo 1967. Since 1975 he is head of the Methods Unit, Central Bureau of Statistics of Norway.

Rec'd: FEB 15 1989
Order No.:
Price: Free
Acc. No.: Central Bureau of
Statistics (Norway)

BOREAL INSTITUTE
LIBRARY

these registers in order to make them useful for statistical purposes. As a result, a considerable amount of statistics is today based on registers. For many statistical purposes, however, the information in the registers is insufficient. Experience has shown that, for instance, registers cannot be substituted for labour force surveys. Similar experiences have been made in connection with other surveys. This does not mean that the registers are of no value when producing statistics. In many cases the information contained in the registers can be used as supplementary information for the purpose of increasing the accuracy of survey results by reducing the sampling variance and the effects of non-response.

The appearance of the first registers gave rise to the hope/expectation that the population census, in due time, would be replaced by the use of registers. In Denmark this has actually happened (Jensen, 1986). In Norway, such expectations have only been fulfilled to a very limited extent. In the 1980 census operation, the registers were used in order to reduce the content of the questionnaire, and at several stages during the collection and preparation of data. However, forms were nevertheless collected from all households.

At present, it is not expected that the 1990 population census will be based on registers alone. It is, however, possible that the 1990 census will involve a combined use of registers and a large sample of households, tentatively expected to be a ten percent sample. In order to evaluate the possibilities and limitations of such an approach, a study based on data from 1980 was undertaken. In the present paper we outline some of the methods we have investigated, and the results and conclusions we have drawn till now from this study.

2. Available registers and their quality

In Norway, a large number of registers is available, which are potentially useful when producing census statistics. The most important are

- (i) *The Central Population Register*, which give up-to-date information on the size of the population, and its distribution by sex, age and geographical regions.
- (ii) *Employer/employee register*, giving information on all employed persons by sex and industry. This register is established for administrative purposes, and the definitions are not identical to those of the census. Furthermore, it is at no time sufficiently up-to-date for all statistical purposes.
- (iii) *Register of income*, being maintained by the tax authorities, which collect information for yearly tax assessment purposes.
- (iv) *Register of education*, containing information on formal education.

In order to evaluate their quality, these registers were used to produce census statistics for 1980, and the results were compared with the results from the ordinary census taken the same year. The comparison was based on a ten percent sample of persons from the census. As each person is identified by an identification number in the registers as well as in the census, linkage was done at the level of the individual.

Comparisons were made for a number of variables (Børke [1]), and the conclusions can be summarized as follows:

(i) For many variables, the discrepancies between the census and the register results were very large, due to differences between the definitions used in the two sources.

(ii) Some census variables, for instance occupation, are not available in any register.

(iii) A population census 1990 should not be based on registers alone.

(iv) Large correlations were observed between the registers and the census.

As large correlations were observed, we studied the possibility and feasibility of reducing the observed register biases through a combined use of registers and census information collected from a sample of the population. A sample selected from the 1980 census was combined with the registers from the same year. The combined results were then compared with the results from the total census. In the next section we shall look at the results from some of these comparisons.

3. Combining registers and samples through post-stratification

As linkage can be done at the level of the individual, and because most variables in the Norwegian census are categorical, we used post-stratification when combining the information from the sample with the census. This technique is simple and easily explained to users of statistics. Besides, it seems highly efficient as shown in Thomsen [6] and Heldal and Spjøtvoll [2]. We illustrate the technique used by showing how a variable with two categories in the census is combined with a similar variable in the register.

Using The Central Population Register and the Employer/employee register, each person can be categorized either as being employed or as being not employed. In the census, a similar categorization is made, but the definition differs from that used in the registers. Our aim is to take a sample from the census, and combine this with the registers in order to estimate the total number of employed persons according to the definition used in the census.

For each person in the sample, we now have the labour force status in the census and in the register, and the following consistency table can be produced:

Table A:

	A	$\bar{A}$	
A	M_{11}	M_{12}	$M_{1.}$
$\bar{A}$	M_{21}	M_{22}	$M_{2.}$
	$M_{.1}$	$M_{.2}$	

M_{11} denotes the number of persons in the sample, which are employed according to both the census and the register.

M_{12} denotes the number of persons in the sample, which are employed according to the census, but not employed in terms of the register.

M_{21} denotes the number of persons in the sample, which are not employed according to the census, but employed according to the register.

M_{22} denotes the number of persons in the sample, which are not employed according to both the census and the register.

In addition to the results in table A, we have the following results from the register:

Table B. The population grouped according to the register:

A	$\bar{A}$	Total
N_1	N_2	N

N_1 denotes the total number of employed persons.

N_2 denotes the total number of unemployed persons.

Based on tables A and B, it is now possible to formulate the following simple estimator for the total number of employed persons according to the census.

$$\hat{E} = \frac{M_{11}}{M_{.1}} N_1 + \frac{M_{12}}{M_{.2}} N_2. \quad (3.1)$$

The first term estimates the number of persons that are employed according to both sources, while the second term estimates those who are not employed according to the register, but employed according to the census.

The estimator, $\hat{E}$, is well known in the sampling literature, as it is a post-stratified estimator. The following points are worth noting:

(i) The estimator $\hat{E}$ has the same expected value as the inflated sample mean. This implies that the definition of employment corresponds to the definition in the census. The statistician has in other words full control over classifications and definitions, which is rarely the case when the register is used as the only source.

(ii) The estimator $\hat{E}$ is unbiased, but it has a sampling variance. It can be shown that

$$\text{var}(\hat{E}) = \text{var}(\tilde{y})(1 - K), \quad (3.2)$$

where $\text{var}(\tilde{y})$ is the sampling variance of the inflated sample mean, and

$$K = \left(\frac{M_{11}}{M_{.1}} - \frac{M_{21}}{M_{.2}} \right)^2.$$

In other words, the combined estimator $\hat{E}$ removes the bias in the register, but it has a sampling variance given by (3.2). For many variables we have found that $K = \frac{1}{2}$, which means that $\hat{E}$ has approximately the same variance as a sample, which is twice the size of the sample actually taken. In the following tables we give examples of how the "register bias" is reduced through post-stratification. In Tables 1, 2 and 3 a simple random sample of 10 000 persons is selected, and linked with the register.

It is seen from Table 1 that the proportion of economically active males is underestimated in the register. Further study reveals that this is primarily due to differences in the definitions used together with the reporting system applied

Table 1

Proportion of economically active males estimated: From the census, the register, and the combined estimator $\hat{E}$

	Post-stratified				
	Census	Register	estimator	(1)–(3)	(1)–(2)
Proportion employed	58.8	57.7	59.3	–0.5	1.1

Table 2

Number of economically active by industry estimated from the census, the register, and the post-stratified estimator, $\hat{E}$

	Census	Register	Post-stratified
Agriculture, forestry and fishing	155 800	141 630	163 360
Mining and quarrying	7 950	8 820	7 780
Manufacturing	407 290	405 650	412 390
Electricity, gas and water	17 280	18 070	19 016
Construction	155 550	131 150	167 600
Wholesale and retail trade, and restaurants and hotels	354 080	326 610	363 140
Transport and communication	173 650	144 460	175 750
Financing, insurance, etc.	103 990	99 350	105 610
Community, social and personal services	618 940	588 700	618 100

Table 3

Number of economically active persons by number of working hours estimated from the census, the register, and the post-stratified estimator $\hat{E}$

Number of working hours per year	Census	Register	Post-stratified estimator
100–499 hours	301 290	189 850	302 501
500–999 hours	238 450	238 800	238 470
1 000–1299 hours	166 710	172 750	169 490
1 300 + hours	1 203 480	1 352 930	1 121 770

when maintaining the register. It is also seen that most of the bias is removed through post-stratification.

Table 2 shows that the register underestimates the number of employed person for most industries. The combined estimator is not exactly equal to the census result, the deviations from the census results being generally smaller than those of the register.

From Table 3 it can be seen that the number of persons working 100–499 hours per year is seriously underestimated in the register. This bias is substantially reduced by using the post-stratified estimator $\hat{E}$.

Many other examples could be presented, but Tables 1–3 clearly demonstrate how the register bias can be substantially reduced by collecting census information from a sample of 10 000 persons, and combining this information with the register.

The method can be generalized by introducing more post-stratification variables, like age, income etc. In the tables given in the next section several post-stratification variables are used.

4. Estimation for small areas

4.1. Synthetic estimates

One of the main reasons for taking censuses is to produce statistics for small areas. It is therefore important to investigate to what extent it is possible to produce statistics for small areas through a combination of samples and registers.

Experiences have already shown that unless the sample size is very large, post-stratification as used in Section 3 is not sufficient. Fortunately, a number of techniques for estimation for small areas have been developed during the last ten years (Purcell and Kish [5], Lundstrøm [4]). Without going into details, we shall outline how these methods can be combined with post-stratification in order to reduce register biases with regard to small area data.

When producing small area data, a post-stratified estimator like (3.1) can be calculated for each area. We shall call this estimate *the direct estimator*. It is unbiased, but due to the sample size it has a considerable sampling variance. In cases where the consistency between the register and the sample is the same across all small areas, or some of them, the coefficients $M_{11}/M_{.1}$ and $M_{12}/M_{.2}$ can be estimated from the whole sample, or a larger part of it. Such an estimator is often called a *synthetic estimator*. Its variance is usually considerably smaller than the variance of the direct estimator, but it may be biased if the consistency between the register and the sample actually varies between the small areas.

One of the main difficulties involved when using synthetic estimators is that no well-established measure of accuracy is developed. In order to evaluate the feasibility of the synthetic estimator, the previously mentioned ten percent linked data set was used to produce the following tables.

From Table 4 it is seen that a synthetic estimator based on a sample of 10 000 persons can substantially reduce the discrepancies between the register and the census. On average, the absolute value of the difference between the register result and the census result is 2.0, while it is 0.5 for the difference between the census result and the synthetic estimate.

For the variable studied, it seems reasonable to conclude that a sample of 10 000 persons, combined with registers, is sufficient to produce census statistics at this regional level with sufficiently good quality.

Using the same synthetic estimator and the same sample size, estimates of the proportion of economically active males were calculated for Norway's 450 muni-

Table 4

Proportion of economically active males by fylke ¹. Estimated from the census, the register and by a synthetic estimator ². Percentages

Fylke	Census	Register	Synthetic estimator
Østfold	58.7	57.5	59.2
Akershus	61.5	60.4	60.9
Oslo	62.8	61.7	64.4
Hedmark	57.7	57.1	57.6
Oppland	58.0	55.8	57.3
Buskerud	58.7	57.8	58.4
Vestfold	57.5	55.4	57.7
Telemark	56.6	56.0	56.7
Aust-Agder	55.1	51.8	54.7
Vest-Agder	58.7	56.8	58.2
Rogaland	59.5	57.5	59.6
Hordaland	58.2	56.0	57.8
Sogn og Fjordane	57.0	53.7	56.0
Møre og Romsdal	58.2	54.2	57.7
Sør-Trøndelag	59.2	57.3	59.1
Nord-Trøndelag	56.8	54.3	56.2
Nordland	56.7	53.8	56.8
Troms	57.1	53.9	57.2
Finnmark	55.5	54.3	56.6

¹ Norway is divided into 19 administrative areas called fylker. They vary in size from approximately 80 000 to 450 000.

² This synthetic estimator is using age and income as post-stratifying variables in addition to the register value. Also the coefficients have been slightly smoothed by a log-linear fitting. The size of the sample is 10 000.

cipalities. Again the conclusion was that they were of satisfactory quality when compared with the actual census results. When the result in Table 4 is broken down by occupation or industry, the conclusion is, however, that a sample of 10 000 is not sufficient. Exactly how large a sample is needed is still under investigation. The relationship between the sample size and the accuracy of a synthetic estimator is complex, and not yet known very well.

In Table 5 synthetic estimators and census results are shown for the proportion of men engaged within primary industries by fylke. Two synthetic estimators are given, one based on a sample of 10 000 persons and one based on 100 000 persons.

Two findings from Table 5 should be noted:

(i) The relative discrepancies between the census results and the synthetic estimates based on 10 000 persons are generally too high for many practical purposes.

(ii) The last two columns show that accuracy increases with sample size. On average, the absolute value of the difference between the census result and the synthetic estimator decreases from 0.8 to 0.5 when the sample is increased from

Table 5

Proportion of men, 16–74 years of age, employed in primary industries by fylke. Census results and synthetic estimators. Percentages

Fylke	(1) Census	(2) Register	(3) Synthetic estimate 10000	(4) Synthetic estimate 100000	(1)–(3)	(1)–(4)
Østfold	5.3	5.7	5.6	5.5	0.3	0.2
Akershus	3.6	3.6	3.8	3.7	0.2	0.1
Oslo	0.3	0.2	1.5	0.9	1.2	0.6
Hedmark	12.3	13.3	10.9	11.9	1.4	0.4
Oppland	13.4	13.3	11.5	12.2	1.9	1.2
Buskerud	5.5	5.9	5.5	5.2	0.0	0.3
Vestfold	4.7	5.1	5.1	4.7	0.4	0.0
Telemark	6.2	7.1	6.1	6.4	0.1	0.2
Aust-Agder	6.8	7.0	6.4	6.3	0.4	0.5
Vest-Agder	5.3	6.5	5.8	5.9	0.5	0.6
Rogaland	8.3	8.9	8.2	8.0	0.1	0.3
Hordaland *	9.5	10.4	9.0	9.7	0.5	0.2
Sogn og Fjordane	16.6	17.7	14.8	15.6	1.8	1.0
Møre og Romsdal	11.7	12.0	10.9	10.9	0.8	0.8
Sør-Trøndelag	8.6	9.4	8.3	8.6	0.3	0.0
Nord-Trøndelag	16.7	16.6	14.1	14.6	2.6	2.1
Nordland	11.2	12.9	11.4	11.4	0.2	0.2
Troms	10.9	12.8	11.5	11.2	0.6	0.3
Finnmark	11.2	14.8	13.0	12.6	1.8	1.4

* Excluding Bergen

10 000 to 100 000. The relationship between the sample size and the accuracy of the synthetic estimator is very “falt”, in the sense that a substantial increase in sample size is necessary to improve the estimate. This has also been observed for other variables, and is an important aspect to have in mind when the size of the sample has to be determined.

4.2. Combined estimators

As mentioned previously, the synthetic estimator is biased, but has a small variance, while the direct estimator is unbiased, and has a typically large variance. In order to find a good estimator for a small area, one often combines the direct and synthetic estimators in the following way:

$$\hat{p}_k = c\hat{p}_1 + (1 - c)\hat{p},$$

where $\hat{p}_1$ is the direct estimator, and $\hat{p}$ is the synthetic estimator. The problem is now to determine c in such a way that the mean square error of $\hat{p}_k$ is minimized. Several approaches are available. In most of our work we have used empirical Bayesian methods. Till now we have very limited experience concerning combined estimators. Based on assumptions concerning the variation between the

small areas, we have developed combined estimators, which are more accurate than the synthetic ones in many cases. In a few cases, however, when the assumption are not satisfied, one runs the risk of a worse performance.

Another difficulty is that combined estimators are much more complex than the synthetic ones. This is a serious problem when the aim is to produce large numbers of tables for several small areas of the country. For each variable, for each cross-classification, and for each small area, an "optimal" estimator may be found. Estimators, which are "optimal" for one set of small areas may not add up to the "optimal" estimator for the larger area. At present we have no good solution to these problems.

5. Final remarks

We can not at present give a simple and final answer to the question raised in the title of this paper. Much more work has to be done. Based on the result achieved till now, it is decided to investigate to what extent it is possible to produce census statistics through a combination of registers with census information from 10 percent of the population. The costs of such a solution is estimated to be 55 percent of those of a traditional census. It is, however, clear that not all statistics usually produced from a census can be produced through a 'combined solution'. For variables characterized by "low" correlations with all register variables, tables must be based on the 10 percent sample exclusively. For other variables, the accuracy and levels of detail are primarily determined by their correlation with variables in existing registers, and by the variation between small areas. At present, it is not possible to state what proportion of the usually produced tables can be obtained through a combined use of registers and a sample.

Even if the final conclusion will be to implement a traditional census in 1990, the efforts to find a "combined solution" are a useful exercise. More insight into the possibilities and limitations of register linkage, and better methods for small area estimation are essential for a more efficient production of statistics in many areas other than population censuses.

References

- [1] S. Børke, Tilleggsundersøkelsen til folke- og bolig-tellingen 1980, Rapporter 84/22, Statistisk Sentralbyrå, Oslo (1984) In Norwegian.
- [2] J. Heldal and E. Spjøtvoll, Statistical calibration of categorical variables with applications to combination of registers and surveys, Contributed Paper, ISI's 45, Session, Amsterdam (1985).
- [3] P. Jensen and L. Thygesen, Linkage of records on object of different kinds, Methodological problems and practical experience, Invited Paper, ISI's 45, Session, Amsterdam (1985).
- [4] S. Lundstrøm, Estimation for small domains: Two studies using combined data from censuses, surveys and registers. Statistik Tidsskrift, Stockholm (1984).
- [5] N.J. Purcell and L. Kish, Postcensal estimates for local areas, International Statistical Review 48 (1980) 3-18.
- [6] I. Thomsen, The use of Markov chain models in sampling from finite populations, Scand. J. Statist. 8 (1981) 1-9.

99999 Pam:31: (*58)
HEL

HELDAL, J. et al.
Census statistics through
combined use of surveys and
registers?

Borrower's Name	Date Due

99999 Pam:31: (*58)
HEL

HELDAL, J. et al.
Census statistics through
combined use of surveys and
registers?

Boreal Institute for Northern
Studies Library
CW 401 Bio Sci Bldg
The University of Alberta
Edmonton, AB Canada T6G 2E9



Tidligere utkommet ISSN 0800-7500:

Previously issued:

- No. 7. Nils Petter Gleditsch, Olav Bjerkholt and Ådne Cappelen: Conversion: Global, National and Local Effects. A Case Study of Norway ISBN 82-537-2180-3
- No. 8. Helge Brunborg: Myter om befolkningsutviklingen ISBN 82-537-2181-1
- No. 9. Johan Heldal and Torbjørn Østdal: Synoptic Monitoring of Water Quality and Water Resources. A Suggestion on Population and Sampling Approaches ISBN 82-537-2207-9
- No. 10. Turid Noack og Lars Østby: Færre barn — Årsaker og konsekvenser ISBN 82-537-2221-4
- No. 11. Erik Biørn, Morten Jensen og Vidar Knudsen: Korttidsdynamikk i norsk økonomi belyst ved kvartalsmodellen KVARTS ISBN 82-537-2240-0
- No. 12. Ådne Cappelen, Nils Petter Gleditsch og Olav Bjerkholt: Military Spending and Economic Growth in the OECD Countries ISBN 82-537-2247-8
- No. 13. John Dagsvik, Boyan Jovanovic and Andrea Shepard: A Foundation for Three Popular Assumptions in Job-Matching Models ISBN 82-537-2283-4
- No. 14. Julie Aslaksen og Olav Bjerkholt: Certainty Equivalence Procedures in Decision-Making under Uncertainty: An Empirical Application ISBN 82-537-2329-6
- No. 15. Olav Bjerkholt og Sigurd Tveitereid: The Use of the MSG-MODEL in Preparing a «Perspective Analysis 1980-2000» for the Norwegian Economy ISBN 82-537-2331-8
- No. 16. Per Schreiner og Knut Arild Larsen: On the Introduction and Application of the MSG-Model in the Norwegian Planning System ISBN 82-537-2332-6
- No. 17. Svein Longva, Lorents Lorentsen og Øystein Olsen: The Multi-Sectoral Growth Model MSG-4 Formal Structure and Empirical Characteristics ISBN 82-537-2333-4
- No. 18. Erik Biørn, Morten Jensen og Morten Reymert: KVARTS — A Quarterly Model of the Norwegian Economy ISBN 82-537-2481-0
- No. 19. Petter Jakob Bjerve: Ragnar Frisch og Økosirksystemet DEL 1 og DEL 2 ISBN 82-537-2493-4
- No. 20. Ådne Cappelen og Knut Mourn: En presentasjon av MODAG-modellenes struktur og egenskaper ISBN 82-537-2522-1
- No. 21. Knut H. Alfsen og Solveig Glomsrød: Future Emissions to Air in Norway: Forecasts Based on the Macroeconomic Model MSG-4E ISBN 82-537-2530-2
- No. 22. Lars Mathiesen, Kjell Roland og Knut Thonstad: The European Natural Gas Market: Degrees of Market Power on the Selling Side ISBN 82-537-2541-8
- No. 23. John K. Dagsvik, Lorents Lorentsen, Øystein Olsen og Steinar Strøm: Residential Demand for Natural Gas: A Dynamic Discrete - Continuous Choice Approach ISBN 82-537-2542-6
- No. 24. Ådne Cappelen og Svein Longva: MODAG A: A Medium-Term Macroeconomic Model of the Norwegian Economy ISBN 82-537-2553-1
- No. 25. Roar Bergan: The Effects of Fiscal Policy in MODAG A ISBN 82-537-2555-8
- No. 26. Nils Martin Stølen: Analysing the Effects of Changes in Wages and Exchange Rates with MODAG A ISBN 82-537-2556-6
- No. 27. Ib Thomsen, Dinke Tesfu og David A. Binder: Estimation of Design Effects and Intraclass Correlations when using Outdated Measures of Size ISBN 82-537-2557-4
- No. 28. Gunvor Iversen, Arne S. Andersen, Anne Lise Ellingsæter og Arvid Hansen: Levekår i Norge: Ulikhet, stabilitet og helhet ISBN 82-537-2561-2